Computer Engineering and Applications Vol. 4, No. 3, September 2015

Data Mining Applications in Big Data

Lidong Wang[1], Guanghui Wang[2]

[1]*Department of Engineering Technology, Mississippi Valley State University, USA*
[2]*State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, China*
lwang22@students.tntech.edu, wanggh@cams.cma.gov.cn
*Corresponding author

ABSTRACT

Data mining is a process of extracting hidden, unknown, but potentially useful information from massive data. Big Data has great impacts on scientific discoveries and value creation. This paper introduces methods in data mining and technologies in Big Data. Challenges of data mining and data mining with big data are discussed. Some technology progress of data mining and data mining with big data are also presented.

Keywords: Big Data, Data Mining, Big Data Analytics, Networks, Grid, Distributed Computing, Stream mining, Web Mining, Text Mining, Information Security.

1. INTRODUCTION

Data mining is a technique for discovering interesting patterns as well as descriptive and understandable models from large scale data. Data mining can be used to find correlations or patterns among dozens of fields in large relational database [1]. Data mining is also the process of discovering or finding some new, valid, understandable, and potentially useful forms of data. Cloud data mining (CDM) is a very tedious process that requires a special infrastructure based on application of new storage technologies, handling, and processing. Big Data/Hadoop is the latest hype in the field of data processing. Through the integration of in-depth analysis of data (data mining) and cloud computing, solutions accessing data mining services every time and everywhere and from various platforms and devices will be made possible [2].

Platform-as-a-service (PaaS) is one of main service models of cloud computing. The PaaS service model stands for the libraries (e.g. *R* library optimized for parallel processing), data mining algorithms, and other services. The benefits of using cloud computing in data mining (DM) are as follows [3]:

- Cost savings – lower operational costs.
- Investment – lower primary investments.
- Faster deployment.
- Easier maintenance – most upgrades and patches are done by the cloud provider.
- Flexibility - ability to add new businesses, spin up new services, and respond to customer

needs.

- Scalability – easier to handle peaks anywhere access and single environment to manage user accounts and credentials across many devices.

Streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge. Data stream can be from sensor networks, measurements in network monitoring and traffic management, click-streams in web exploring, manufacturing processes, and twitter posts, etc. [4]. Data stream mining studies methods and algorithms for extracting knowledge from volatile streaming data. Streaming data needs fully automated preprocessing methods. Preprocessing models need to be able to update themselves automatically along with evolving data. Furthermore, all updates of preprocessing procedures need to be synchronized with the subsequent predictive models. Therefore, not only models, but also the procedure itself needs to be fully automated. Only a small subset of stream-based selective sampling algorithms is suited for non-stationary environments [5]. Streaming data processing is also a method of big data processing. Streaming data is temporal data in nature. Streaming data may also include spatial characteristics [6].

Big data mining is the capability of extracting useful information from these large datasets or streams of data, which was not possible before due to data's volume, variability, and velocity [7]. Big data is a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques. Big data technologies have great impacts on scientific discoveries and value creation [8, 9, 10]. Structured (numerical) and unstructured (textual) are two main types of data forms in big data. Their characteristics and uses are listed in Table 1 [11].

TABLE 1.
Characteristics and uses of structured and unstructured data

Characteristics	Structured Data	Unstructured Data
Variety	Instrumented or known sources; typically row s and columns of numbers	Unknown sources; typically critical in stances of specific words in a context of interest
Volume	Large and fast growing; continually aggregating amassed data to assess decisions	May or may not be large; seeks to isolate specific information to make decisions
Velocity	Real time and/or archival; used for operational efficiency	Not as fast or archival; used for strategic decision making
Velocity	Data are auditable; sources can be validated	Multiple sources must be used to triangulate validity

Web mining can be divided into three different types. They are: Web usage mining, web content mining, and Web structure mining. Web usage mining is a process of extracting useful information from server logs, i.e. user's history. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Web content mining aims to discovering useful information or knowledge from web page contents rather than hyperlinks and goes beyond using keywords in a search engine. Web content consists of information such as unstructured free text, image, audio, video, metadata, and hyperlink [12].

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

The real existing problem is most data mining methods do not work well with big data. There are a lot of challenges when data mining methods apply to Big Data analytics. The objective of this paper is to identify what data mining methods can be used in big data and present the improvements or novelties of these methods through introducing the technology progress of data mining with big data. This paper introduces data mining, data mining with big data, and the challenges and technology progress of data mining with big data. The challenges presented in this paper partly indicate the gap/problem from previous research work as well as some future work. Therefore, this paper will present some significant value of data mining applications in big data. The organization of this paper is as follows: the next section introduces methods of data mining and Big Data; Section 3 discusses challenges of data mining and data mining with big data; Section 4 presents technology progress of data mining and data mining with big data; and the final section is conclusions.

2. METHODS OF DATA MINING AND BIG DATA

Data mining is a set of techniques for extracting valuable information (patterns) from data. It includes clustering analysis, classification, regression, and association rule learning, etc. [13]. For example, cluster analysis is used to differentiate objects with particular features and divide them into some categories (clusters) according to these features. It is an unsupervised study method without training data. Clustering can be considered the most important unsupervised learning problem [1, 14]. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. Several major kinds of classification algorithms in data mining are decision tree, k-nearest neighbor (KNN) classifier, Naive Bayes, Apriori and AdaBoost [1]. Regression analysis identifies dependence relationships among variables hidden by randomness [14].

KNN classifiers are a kind of nonparametric method for classifying data objects based on their *k* closest training data objects in the data space. The KNN classifiers do not construct any classifier model explicitly; instead they keep all training data in memory. Hence they are not amenable to big data applications [15].

Data mining services exploit and are built on top of a cloud infrastructure and other most prominent large data processing technologies to offer functionalities such as high performance full text search, data indexing, classification and clustering, directed data filtering and fusion, and meaningful data aggregation. Advanced text mining techniques such as named entity recognition, relation extraction, and opinion mining help extract valuable semantic information from unstructured texts. Intelligent data mining techniques that are being used include local pattern mining, similarity learning, and graph mining [16].

In streaming data mining, Very Fast Decision Tree (VFDT) is a streaming data classifier which starts with only the root node, sorts training data to leaf nodes, and splits the leaf nodes that meet the splitting criteria on-the-fly. It can be successfully applied to stream data, but it has some restrictions to apply big data because the quality measures like the information gain for splitting attributes are evaluated over (yet big) data subsets [15].

A way of speeding up the mining of streaming learners is to distribute the training process onto several machines. Hadoop is such a programming model and

software framework. Apache *S*4 is a platform for processing continuous data streams. *S*4 applications are designed for combining streams and processing elements in real time [4].

Streaming data processing and mining have been deploying in real-world systems such as InforSphere Streams (IBM), Rapidminer Streams Plugin, StreamBase, MOA, AnduIN [6]. SAMOA is a new upcoming software project for distributed stream mining that will combine *S*4 and Storm with MOA [7].

There are a lot of Big Data technologies. Massive parallel-processing (MPP), Hadoop, NoSQL, and MPP databases, etc. have been used to support Big Data [17]. Table 2 [18] compares a number of Big Data technologies. The table highlights the different types of systems and their comparative strengths and weaknesses.

TABLE 2.
Comparison of Big Data Technologies

	In-Memory Database	MPP Database	Big Data Appliance	Hadoop	NoSQL Database
Consistent	W	W	W	P	P
Available	W	W	W	P	P
Fault tolerant	W	W	P	W	W
Suitable for real-time transactions	W	W	W	F	F
Suitable for analytics	P	P	W	W	F
Suitable for extremely big data	F	P	P	W	W
Suitable for unstructured data	F	F	P	W	W

W : Meets widely held expectations.
P : Potentially meets widely held expectations.
F : Fails to meet widely held expectations

In big data mining and analysis, some tools and popular open source initiatives are as follows [4, 14]:

- *Apache Mahout*: Scalable machine learning and data mining software based mainly on Hadoop. It has implementations of clustering, classification, collaborative filtering, and frequent pattern mining.
- *MOA*: Stream data mining software to perform data mining in real time. It has implementations of clustering, classification, regression, frequent item set mining, and frequent graph mining.
- *R*: open source programming language and software environment designed for statistical computing, data mining/analysis, and visualization.
- *GraphLab*: high-level graph-parallel system built without using MapReduce.
- *Excel*: It provides powerful data processing and statistical analysis capabilities.
- *Rapid-I Rapidminer*: Rapidminer is open source software used for data mining, machine learning, and predictive analysis. Data mining and machine learning programs provided by RapidMiner include Extract, Transform, and Load (ETL); data pre-processing and visualization; modeling, evaluation, and deployment.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

- *KNIME*: Konstanz information miner (KNIME) is a user-friendly, intelligent, and open-source rich data integration, data processing, data analysis, and data mining platform.
- *Weka/Pentaho*: Weka is a free and open-source machine learning and data mining software written in Java. Pentaho includes a web server platform and several tools to support reporting, analysis, charting, data integration, and data mining, etc.

3. CHALLENGES OF DATA MINING AND DATA MINING WITH BIG DATA

Protecting privacy and confidentiality, stream preprocessing, timing and availability of information, and relational stream mining, etc. are challenges. Challenges of data stream processing and mining lie in the changing nature of streaming data. Therefore, identifying trends, patterns, and changes in the underlying processes generating data is important [5, 6].

Data streams pose challenges for data mining. First, algorithms must make use of limited resources (time and memory). Second, they must deal with data whose nature or distribution changes over time [4]. Unique challenges associated with designing distributed mining systems are: online adaptation to incoming data characteristics, online processing of large amounts of heterogeneous data, limited data access and communication capabilities between distributed learners, etc. [19].

The general MapReduce mode is not suitable for data mining. First of all, MapReduce is lack of overall. The lack of data sharing between the tasks nodes in Hadoop, such as shared memory. Secondly, the Hadoop distributed file system (HDFS) does not allow random write operation. Massive data once written into the HDFS only can be added or deleted. Thirdly, the task has a short life cycle. Finally, MapReduce may not be well suited for complex algorithms that have an iterative nature [20].

Big data mining is more challenging compared with traditional data mining algorithms. Taking clustering as an example, a natural way of clustering big data is to extend existing methods (such as hierarchical clustering, K-Mean, and Fuzzy C-Mean) so that they can cope with the huge workloads. Most extensions usually rely on analyzing a certain amount of samples of big data, and vary in how the sample-based results are used to derive a partition for the overall data. Clustering big data is also developing to distributed and parallel implementation [13]. Lack of computational performance and storage capacity were identified as the main obstacles of cloud computing in data mining with big data [3].

Big data mining has challenges in data accessing and computing procedures. Big data are often stored at different locations. While typical data mining algorithms require all data to be loaded into the main memory, moving data across different locations is expensive. Big data mining challenges and difficulties in algorithm designs are raised by the big data volumes, distributed data distributions, and by complex and dynamic data characteristics [21]. The challenges of big data mining algorithms are listed as follows [21]:

- *Local learning and model fusion for multiple information sources*: A big data mining system has to enable an information exchange and fusion mechanism

to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal.

- *Mining from sparse, uncertain, and incomplete data*: Sparse data cannot be used to draw reliable conclusions. Common approaches are to employ dimension reduction or feature selection to reduce the data dimensions or to carefully include additional samples, such as generic unsupervised learning methods in data mining. For uncertain data, each data item is represented as some sample distributions. Common solutions are to take the data distributions into consideration to estimate model parameters. Most data mining algorithms can handle incomplete or missing data. Imputing missing values is a method of producing improved models.
- *Mining complex and dynamic data*: Currently, there is no acknowledged effective and efficient data model to handle big data complexity (structured, unstructured, and semi-structured).

Mining big data streams faces three principal challenges: *volume*, *velocity*, and *volatility*. Volume and velocity require a high volume of data to be processed in limited time. Volatility corresponds to a dynamic environment with ever-changing patterns. Old data is of limited use. This is due to change that can affect the induced data mining models in multiple ways: change of the target variable, change in the available feature information, and drift [5, 6]. Mining heterogeneous information networks is a promising research frontier in big data mining. Existing data mining techniques face great difficulties when they are required to handle big data. Key issues and challenges are heterogeneity, volume, speed, accuracy, garbage mining, and crisis in trust and privacy [22].

4. TECHNOLOGY PROGRESS OF DATA MINING AND DATA MINING WITH BIG DATA

A general framework for distributed data mining was proposed and an efficient online learning algorithm was developed. The proposed learning algorithms can optimize the prediction accuracy while requiring significantly less information exchange and computational complexity [19].

Outlier detection is important in data mining. Various methods for outlier detection have been developed particularly for dealing with numerical data. A two-phase algorithm for detecting outliers in categorical data was proposed based on a novel definition of outliers. In the first phase, this algorithm explores a clustering of the given data, followed by the ranking phase for determining the set of most likely outliers. The proposed algorithm is expected to perform better as it can identify different types of outliers, employing two independent ranking schemes based on the attribute value frequencies and the inherent clustering structure in the given data [23].

Privacy and security concerns restrict the sharing or centralization of data. Privacy-preserving data mining has emerged as an effective method to solve this problem. Distributed solutions have been proposed that can preserve privacy while still enabling data mining. However, while perturbation based solutions do not provide stringent privacy, cryptographic solutions are too inefficient and infeasible to enable truly large scale analytics for big data. A solution that uses both randomization and cryptographic techniques was proposed to provide improved

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

efficiency and security for several decision tree-based learning tasks. The proposed approach is based on random decision trees (RDT). The same code of RDT can be used for multiple data mining tasks: classification, regression, ranking, and multiple classifications. RDT is also excellent in privacy preserving distributed data mining [24].

Density estimation is the ubiquitous base modelling mechanism employed for many tasks including clustering, classification, anomaly detection and information retrieval. Commonly used density estimation methods such as kernel density estimator and *k*-nearest neighbor density estimator have high time and space complexities which render them inapplicable in problems with big data. A density estimation method was proposed for dealing with millions of data easily and quickly. An asymptotic analysis of the new density estimator was provided and the generality of the method was verified by replacing existing density estimators with the new one in three current density-based algorithms, namely DBSCAN, LOF and Bayesian classifiers, representing three different data mining tasks of clustering, anomaly detection and classification [25].

Data stream mining has shown the potential to be beneficial for clinical practice. By using data stream diagnosis for prognosis and spell detection, physicians could make faster and more accurate decisions. Data mining and Big Data analytics are helping to realize the goals of diagnosing, treating, helping, and healing all patients in need of healthcare. In order to handle the continuous stream of data, an algorithm that can handle high-throughput data will be necessary. Very Fast Decision Tree (VFDT) was used for this purpose. VFDT has many advantages over other methods (e.g., rule based, neural networks, other decision trees, Bayesian networks). It can make prediction both diagnostically and prognostically and handle a changing non-static dataset [26].

A classification method which can handle big data with both categorical and numerical attributes was proposed. The method partitions the numerical data space into a grid structure and makes each grid cell maintain probability distributions of both categorical and numerical attributes. Using the probability distributions of the *k*-nearest neighbor cells as well as the home cell, the class label of query data is determined by Bayesian inference [15].

Frequent itemset mining (FIM) is a method to extract knowledge from data. FIM tries to extract information from databases based on frequently occurring events according to a user given minimum frequency threshold. The combinatorial explosion of FIM methods has become problematic when they are applied to big data. Two algorithms that exploit the MapReduce framework were proposed to deal with two aspects of the challenges of FIM for mining big data: (1) Dist-Eclat is a MapReduce implementation of the well-known Eclat algorithm, optimized for speed in case a specific encoding of the data fits into memory. (2) BigFIM is optimized to deal with truly big data by using a hybrid algorithm, combining principles from both Apriori and Eclat, also on MapReduce. The experiments showed that the proposed methods outperformed state-of-the-art FIM methods on big data using MapReduce [27].

Heterogeneous mixture learning, the most advanced heterogeneous mixture data analysis technology, was developed by *NEC* Corporation in Japan. The heterogeneous mixture learning technology is an advanced technology used in big data analysis. As the big data analysis increases its importance, heterogeneous

mixture data mining technology is also expected to play a significant role in the market. The range of application of heterogeneous mixture learning will be expanded broader than ever in the future [28].

In order to mine big data in real-world applications, it is necessary to efficiently identify a fixed number of relevant features for building accurate prediction models in the online learning process. A new research problem of online feature selection (OFS) was investigated, which aimed to select a fixed number of features for prediction by an online learning fashion. A novel OFS algorithm was presented to solve the learning task, and theoretical analysis on the mistake bound of the proposed OFS algorithm was offered. Results showed the proposed algorithms were fairly effective for feature selection tasks of online applications, and significantly more efficient and scalable than some state-of-the-art batch feature selection technique [29].

Two phases was employed for sentiment analysis: preprocessing using natural language tool kit (NLTK) and data mining using Mahout. Mahout is an open source machine learning library from Apache for big data analysis. The sentiment mining of Twitter data was implemented using Mahout. MapReduce framework was incorporated so as to implement the work in a distributed environment using Mahout [30]. Pre-processing data helps in dimensionality reduction, thereby eliminating a lot of unnecessary features from being handled and in some cases making In-Memory processing of data possible, resulting in a huge reduction of I/O overhead [31].

5. CONCLUSIONS

Data mining can be used to discover hidden, unknown, but useful knowledge from massive, fuzzy, noisy, incomplete, and random data. The KNN classifiers are not amenable to big data applications. VFDT can be applied to stream data, but it has some restrictions to apply big data because the quality measures like the information gain for splitting attributes are evaluated over data subsets.

Big Data analytics requires that distributed mining of data streams should be performed in real-time. Much research is needed in practical and theoretical analysis to provide new methods for distributed data mining with big data streams.

The challenges of big data mining algorithms are: local learning and model fusion for multiple information sources; mining from sparse, uncertain, and incomplete data; and mining complex and dynamic data.

Some technology progress has been made such as the classification method for big data with both categorical and numerical attributes, heterogeneous mixture learning, and the online feature selection (OFS) algorithm, etc. The above challenges about data mining with big data can be further research topics.

REFERENCES

[1] B. Thakur, M. Mann, "Data Mining for Big Data: A Review," *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, No. 5, pp. 469-473, 2014.

[2] R. Vrbić, "Data mining and cloud computing," *Journal of Information Technology & Applications*, Vol. 2, No. 2, pp. 75-87, 2012.

[3] V. Nekvapil, "Cloud computing in data mining – a survey," *Journal of Systems Integration*, No. 1, pp. 12-23, 2015.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[4] A. Bifet, "Mining Big Data in Real Time," *Informatica*, Vol.37, pp. 15–20, 2013.
[5] G. Krempl, I. Zliobaite, D. B. Nski, E. H. Ullermeier, et. al., "Open Challenges for Data Stream Mining Research," *ACM SIGKDD Explorations*, Vol. 16, No. 1, pp. 1-10, 2013.
[6] D.-H. Tran, M. M. Gaber, K.-U. Sattler, "Change detection in streaming data in the era of big data: models and issues," *ACM SIGKDD Explorations*, Vol. 16, No. 1, pp. 30-38, 2014.
[7] W. Fan, A. Bifet, "Mining Big Data: Current Status, and Forecast to the Future," *ACM SIGKDD Explorations*, Vol. 14, No. 2, pp. 1-5, December 2012.
[8] Y. Demchenko, P. Grosso, C. D. Laat, P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," 2013 International Conference on Collaboration Technologies and Systems (CTS), 20-24 May 2013, San Diego, CA, USA, pp. 48-55, 2013.
[9] D.E. O'Leary, "'Big Data', the 'Internet of Things' and the 'Internet of Signs'," *Intelligent Systems in Accounting, Finance and Management*, Vol. 20, pp. 53-65, 2013.
[10] H.V. Jagadish, A. Labrinidis, Y. Papakonstantinou, *et al*., "Big Data and Its Technical Challenges," *Communications of the ACM*, Vol. 57, No. 7, pp. 86-94, 2014.
[11] S. K. Markham, M. Kowolenko, and T. L. Michaelis, "Unstructured Text Analytics to Support New Product Development Decisions," *Research Technology Management*, pp. 30-38, March-April, 2015.
[12] S. B. Boddu, "Eliminate the noisy data from web pages using data mining techniques," Computer Science and Telecommunications, Vol. 38, No. 2, pp. 39-46, 2013.
[13] C.L. P. Chen, C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, Vol. 275, No. 10, pp. 314-347, 2014.
[14] M. Chen, S.-W. Mao, Y.-H. Liu, "Big data: A survey," *Mobile Netw Appl*, Vol. 19, pp. 171-209, 2014.
[15] K. M. Lee, "Grid-based Single Pass Classification for Mixed Big Data," *International Journal of Applied Engineering Research*, Vol. 9, No. 21, pp. 8737-8746, 2014.
[16] N. Karacapilidis, M. Tzagarakis and S. Christodoulou, "On a meaningful exploitation of machine and human reasoning to tackle data-intensive decision making," *Intelligent Decision Technologies*, Vol. 7, pp. 225–236, 2013.
[17] M. Turk, A chart of the big data ecosystem, take 2, 2012. http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/
[18] J. Dean, "Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners," John Wiley & Sons, Inc., 2014.
[19] Y. Zhang, D. Sow, D. Turaga, M. v. d. Schaar, "A Fast Online Learning Algorithm for Distributed Mining of BigData," *ACM SIGMETRICS Performance Evaluation Review*, Vol. 41, No. 4, pp. 90-93, 2014.
[20] S. Londhea, S. Mahajan, "New Approach For Big Data Mining Using MapReduce Techniques," *International Journal of Applied Engineering Research*, Vol. 10, No. 6, pp. 15407-15415, 2015.

[21] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, "Data Mining with Big Data," *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 26, No. 1, pp. 97-107, 2013.
[22] K. Pal, J. Saini, "A Study of Current State of Work and Challenges in Mining Big Data," *International Journal of Advanced Networking Applications*, Special Issue, pp. 73-76, 2014.
[23] N.N.R. R. Suri, M. N. Murty and G. Athithan, "A ranking-based algorithm for detection of outliers in categorical data," *International Journal of Hybrid Intelligent Systems*, Vol. 11, pp. 1–11, 2014.
[24] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, "A Random Decision Tree Framework for Privacy-Preserving Data Mining," *IEEE Transactions on Dependable and Secure Computing*, Vol. 11, No. 5, pp. 399-411, 2014.
[25] K. M. Ting, T. Washio, J. R. Wells, F. T. Liu, S. Aryal, "DEMass: a new density estimator for big data," *Knowledge & Information Systems*, Vol. 35, pp. 493–524, 2013.
[26] M. Herland, T. M Khoshgoftaar and R. Wald, "A review of data mining using big data in health informatics," *Journal of Big Data*, Vol. 1, No. 2, pp. 1-35, 2014.
[27] S. Moens, E. Aksehirli and B. Goethals, "Frequent Itemset Mining for Big Data," 2013 IEEE International Conference on Big Data, 6-9 Oct. 2013, Silicon Valley, CA, USA, PP. 111- 118.
[28] F. Ryohei, M. Satoshi, "The Most Advanced Data Mining of the Big Data Era," *NEC Technical Journal*, Vol.7 No.2, PP. 91-95, 2012.
[29] S. C.H. Hoi, J. Wang, P. Zhao, R. Jin, "Online Feature Selection for Mining Big Data," Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, PP. 93-100, 2012.
[30] U. Jaswant and P.N. Kumar, "Big Data Analytics: A Supervised Approach for Sentiment Classification Using Mahout: An Illustration," *International Journal of Applied Engineering Research*, Vol. 10, No. 5, pp. 13447-13457, 2015.
[31] V.J. Nirmal and D.I.G. Amalarethinam, "Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data," *International Journal of Fuzzy Mathematical Archive*, Vol. 6, No. 2, pp.149-159, 2015.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)